图书情报工作 ChinaXiv合作期刊
第 63 卷 第 20 期 2019 年 10 月

网络舆情衍进指数构建与实证分析*

黄微 朱镇远 许烨婧 孙悦
吉林大学管理学院 长春 130022

摘要：［目的/意义］提出和构建网络舆情衍进指数，以描述网络舆情演化过程中常衍生出新的子话题的现象，对于舆情预警、预测具有重要的理论及实践意义。［方法/过程］以文本聚类结果和文本聚类有效性为依据，提出网络舆情衍进的判别标准和舆情衍进指数的构建过程，并以“教科书老赖”这一事件作为样本数据进行实证分析。［结果/结论］所构建的舆情衍进速率指数可以用于描述舆情衍进。在突发期阶段话题舆情衍进指数最高，此后逐渐下降，这一阶段的舆情衍进最为剧烈，子话题的出现呈现爆发性增长；舆情衍进指数在舆情蔓延期内出现阶梯式下降，此后保持为负值，舆情的子话题开始逐渐减少，舆情内容本身由发散转为收敛；进入消散期后，子话题数量趋于稳定。作为舆情衍进速率的测度和舆情衍进的判别方式，舆情衍进指数为舆情监管和舆情预警提供了全新的角度。

关键词： 舆情衍进 衍进指数 文本聚类 聚类有效性
分类号： G251
DOI：10.13266/j.issn.0252-3116.2019.20.003

1 引言

网络舆情相较于传统媒体而言，具有传播速度快、信息量大、互动性强、准入门槛低等特点[1]。在网络环境中，社会个体能够更加便捷地发布和获取信息，与此伴生的舆情的高不确定性——舆情产生新子话题的概率呈指数性增长，舆情衍进具有较高的时间敏感度——从以下两个方面对政府新媒体平台上的舆情监管提出了挑战：①类似舆情事件产生联动效应，多角度触动舆情受众神经，舆情存在迸发的可能；②相较于原始话题，子话题发生了不可逆转的变化，出现了与原始话题内涵不同的“民粹”式反馈，如“平安高管是老赖”“唐山法院不作为”等报道。此类话题走向难以预测，舆情受众观点存在负向极化风险。政府及舆情相关管控部门对热点舆情的衍进变化应采取高度审慎的态度，从而更好地实现国家“塑造清朗的网络空间”的诉求。同时，将网络舆情衍进这一现象从网络舆情整体演化过程中抽出，构建单独的指数，对于网络舆情衍进的实时监控和快速响应而言，具有重要的理论和现实意义。

目前而言，国内外学者对网络舆情衍进指数的相关研究主要集中于以下 2 个方面：

（1）将舆情衍进视为新要素加入舆情传播模型，以研究舆情传播规律。依据网络衍生舆情传播的不同特性，高宾等将网络衍生舆情分为 5 种类型，对不同的网络衍生舆情模型提出相应的概率计算方法，并对网络衍生舆情的分析流程进行了具体的描述[2]。K. Saito 等对网络舆情传播中的各个节点属性加以分析，从而得出各节点进行舆情传播的概率[3]。D. J. Watts 等的研究表明舆情传播的起点是否为意见领袖并非舆情演进的关键因素[4]。兰月新等以 logistic 模型为基础模型，构建网络舆情衍生效应的数学模型，通过模型平衡点及稳定性研究不同信息异化条件下正面衍生舆情和负面衍生舆情的传播特性[5]。陈福集等的研究在传染病传播模型基础上引入话题衍生率，构建了 SEIRS 网络舆情传播演化模型。通过对传播阈值和平衡点的求解，从理论上分析了话题衍生率对传播态势的影响，并依据数值仿真模拟实验分析了不同因素对网络舆情传播规律的影响[6]。尹熙成等研究了衍生话题与原话题在网络中独立传播并相互影响的过程，得出了衍生

* 本文系国家自然科学基金面上项目“大数据环境下多媒体网络舆情信息的语义识别与危机响应研究”（项目编号：71473101）研究成果之一。
作者简介：黄微（ORCID：0000-0003-0448-9563），教授，博士生导师，E-mail：huangwei@jlu.edu.cn；朱镇远（ORCID：0000-0002-5247-0608），博士研究生；许烨婧（ORCID：0000-0003-1128-2878），博士研究生；孙悦（ORCID：0000-0001-7343-8343），硕士研究生。
收稿日期：2019-03-08 修回日期：2019-05-27 本文起止页码：26-33 本文责任编辑：易飞

话题会使舆情传播过程出现新的高峰点,话题的转发率显著提高,导致话题演化的弛豫时间延长的结论[7]。王丽君等的研究根据网络舆情衍生的共性规律,归纳出3种不同的衍生链结构,并给出了对应的衍生概率算法[8]。总体而言,这一类型研究的侧重点在于研究话题衍生衍进对于舆情整体演化传播过程的影响,对于舆情衍进的判定标准并没有给出较为具体的定义。

(2)针对具体舆情事件建立相关指标体系。靳晓宏等结合以往指标体系构建的相关研究,从5个维度构建了主题事件舆情指标体系,并基于主题事件类指标体系,以食品安全为例,通过层次分析法得出舆情指标的权重,构建了舆情指数[9]。贺恩峰等从传播媒体、传播范围、传播速度、情绪倾向程度及相关度等方面对舆情潜在影响力进行探索,同样利用层次分析法得出了舆情潜在影响力指标体系各因子权重系数[10]。邓尚民等基于AHP和调查法,对高校网络舆情安全评估设计了相应的警源指标和警兆指标,用以构建高校网络舆情安全评估指标体系[11]。这一类型的研究,其侧重点主要在于构建描述舆情演化或进行舆情预警的综合体系以及指标体系中各项指标的权重,对于舆情衍进这一现象,尚未有单独的判别标准和对应指标。

综上所述,目前国内外学者对于舆情衍进的研究,更多是在研究舆情传播的过程中加入舆情衍进这一因素。对于舆情衍进本身量化判别尚存研究空间。各类舆情相关指标的研究,也并未对网络舆情衍进这一现象单独建立指标。针对舆情或网络热点的指数研究,除学术界外,工业界中也存在着相关的商业应用。从全球范围来看,本文所研究的网络舆情衍进指数与谷歌所开发的Google Trend存在一定相似性。Google Trend可根据用户输入的关键词,提供与关键词相关的话题,其底层算法是利用用户搜索关键词的相关度来提供相关话题,与本文所采用的文本聚类及文本聚类有效性的研究思路并不相同,同时其所提供的相关话题也并非全部为该关键词的子话题。若将范围限定在中文,业内目前应用最为广泛的网络指数百度指数聚焦于关键词搜索趋势及搜索用户画像构建这两点上,对于子话题的衍生消亡过程也并没有给出明确的表现。

结合前人在话题发现和舆情传播方面的研究成果,本文试图构建网络舆情衍进速率指标,用以描述网络舆情衍进这一现象,从而为网络舆情监控与预警提供新的视角。本文的研究,在理论层面确定网络环境下舆情衍进标准,构建网络舆情衍进速率的关键指数;在实践方面,以新浪微博"教科书式老赖"这一热点话题作为样本数据进行文本聚类,结合事件本身发展的时间序列对比聚类结果,以验证网络舆情衍进指数的适用性。

2 舆情衍进系数相关理论

2.1 舆情演进与舆情衍进

目前针对舆情演进的研究对于"舆情演进"并无统一的定义,同时,诸多研究中的"演进""演变""演化"等概念并无本质区别[12]。结合前人对于网络舆情演进的研究,本文将网络舆情"演进"和"衍进"的概念区分如下:

(1)所谓网络舆情演进,是指单一网络舆情在时间、空间、规模、议题、热度、受众群体等多个维度上,从发生、发展、高峰、波动到淡化、消亡的整体过程。在这一概念下,针对不同舆情事件具体如何演进,国内学者利用不同的数学模型从多个角度进行了阐述。周昕等以多媒体技术、舆情分析理论、信息传播理论为支撑,对网络舆情传播方式受多媒体技术的影响情况加以揭示,深入剖析传统网络舆情传播模式[13]。黄微等人构建了微博舆情信息老化模型,为微博舆情信息的监测提供计算支持[14]。

(2)所谓网络舆情衍进,在本文的研究范畴中,特指单一网络舆情演进过程中,衍生出新生子话题、子舆情的过程。正如引言部分所述,目前舆情衍进的相关研究多将舆情衍进作为单一要素纳入舆情传播模型中进行考量,对于舆情衍进的具体标准尚缺乏深入的探讨。

2.2 网络舆情衍进指数

根据前文对网络舆情衍进的定义,本文对网络舆情衍进指数(Public Opinion Derivative Index)的定义如下:网络舆情衍进指数,是指单一网络舆情在某一具体时刻下,衍生出新生子话题的速率。舆情衍进,是舆情演进过程中内容丰富度及舆情文本复杂度发生变化的过程。正如引言中所述,现存的网络舆情指标体系中,并未针对网络舆情衍进这一现象构建单独的指标。本文利用文本聚类及聚类有效性所得出的网络舆情衍进指数,以量化的方式对舆情衍进给出了判别标准。相较于以往关注网络舆情演化整体过程的指标体系而言,该指数专注于舆情衍进这一单一要素。相关部门可通过监测网络舆情衍进指数实现舆情衍进预警,从而有针对性地进行舆情管控。

3　网络舆情衍进指数构建

3.1　网络舆情衍进指数构建过程

网络舆情衍进指数的构建过程如下：在获得网络舆情数据后，首先需要对数据进行预处理，这一流程包括了对非标准化数据的标准化处理、对标准化后的纯文本舆情数据进行分词、删除停用词等步骤，并构建舆情词袋空间；其次，利用 TF-IDF 值计算特定词权重，再利用 K-Means 聚类方法对不同初始 K 值下的结果进行文本聚类；最终，对比不同 K 值下文本聚类的聚类有效性结果，确定当前时刻最优的话题数量 K′，作为舆情衍进的判别标准。如图 1 所示：

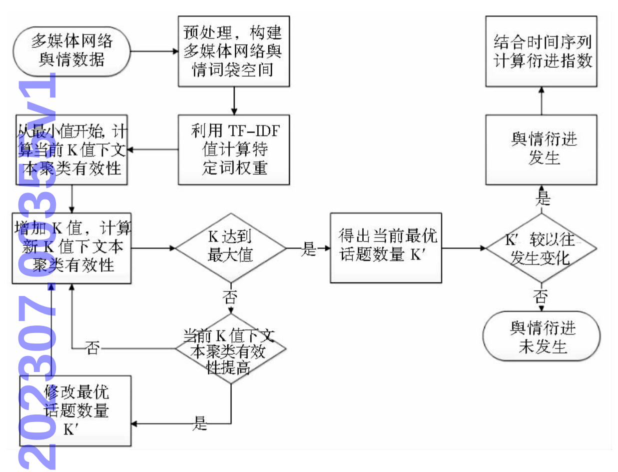


图 1　网络舆情衍进指数构建过程

3.2　文本预处理与构建词袋空间

为了对网络舆情进行文本聚类，首先需要对舆情数据进行预处理，使非标准化舆情数据转化为标准化纯文本数据。此后，需要对文本进行切词，本文采用 jieba 分词工具包于 Python 环境下实现中文文本分词，采用其中的精准分词模式，试图将文本中的句子最精确地分开，以适应文本分析的需求。在分词后，对数据进行去除虚词、代词等停用词处理以提高语料库的信息密度。虽然 jieba 工具包中已经包含了去除停用词功能，但其主要是为其本身的文本分析工具所使用，不利于后续分析流程，因此本文另行采用了包含 1 893 个常用中文停用的停用词表，以去除停用词。在去除停用词后，统计所有文档词集合，针对每个文档构建向量，向量的值即是某一词在该文档中出现的次数。由此，即构建成功了词袋空间 VSM（vector space model）。

3.3　利用 TF-IDF 计算特定词权重

针对已经构建的词袋空间，本文采取 TF-IDF（term frequency-inverse document frequency）方法，将词所出现的次数转化为在语料库中的权值。该方法认为，字词的重要性与它在单一文本中所出现的次数成正比，而与它在语料库中所出现的频率成反比。

该方法中，词频（term frequency，TF）指某一给定词语在某一文件中出现的频率，该数值是对词数（term count）的归一化，具体表达式如下：

$$tf_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}}$$

其中 $n_{i,j}$ 是词 t_i 在文件 d_j 中的出现次数，而分母则是文件 d_j 中所有字词出现次数之和。

逆向文件频率（inverse document frequency，IDF）是一个词语在语料库中普遍重要性的度量。对于某一特定词语的 IDF，可由总文件数除以包含该词文件数目再对所得结果取对数而得：

$$idf_i=\log\frac{|D|}{1+|\{j: t_i\in d_j\}|}$$

其中 $|D|$ 为语料库中文件的总数，$|\{j: t_i\in d_j\}|$ 为包含词语 t_i 的文件个数，当然，如果该词语不在语料库中，会导致被除数为 0，因此一般情况下使用 $1+|\{j: t_i\in d_j\}|$。

最终的 TF-IDF 值为：

$$tfidf_{ij}=tj_{i,j}\times idf_i$$

该值为考虑到了单一词汇在特定文件中所出现的高词频和在整个语料库中的低词频后得出的综合权重值，因而倾向于过滤掉常见词语，而保留重要的高信息量词。

通过计算词袋空间向量后的矩阵，其列为所有文档词的集合，每一行代表一个文档，而向量的值则为该词在整体语料库中和该文本中的权值。

3.4　基于 K-Means 算法的文本聚类

利用计算过 TF-IDF 值的矩阵，可以采用多种方式进行聚类分析。在本文的研究中，我们采用 K-Means 算法进行文本聚类。K-Means 算法是一种经典的基于划分的聚类算法，其基本原理是首先随机选择 K 个文档（在经过 TF-IDF 值计算后，为矩阵中的一个向量）作为初始聚类点，然后根据簇中对象的平均值，将剩余文档归类给最类似的簇，同时更改簇的平均值。如此重复迭代一定次数，直至簇的划分不再改变。具体计算步骤如下所示：

（1）输入语料库随机选取 k 行作为聚类初始中心。

（2）将剩余数据中的一个分配至与之欧式距离（$Dis_2(x,y)=\sqrt{\sum_{j=1}^{d}(x_i-y_i)^2}$）最近的聚簇中。

（3）更新聚簇集合 C 和聚簇均值。

（4）重复以上过程，直至目标函数 $\sum_{i=1}^{n}(\arg\min\|x_i-c_j\|_2^2)$ 收敛。

K-Means 算法由于其简单、高效的特征在文本聚类中得到了广泛的应用。

3.5 基于文本聚类有效性指数的最优话题数量判断

在传统的 K-Means 聚类中，除了初始聚类簇中心的选取外，K 值本身的选取也至关重要。通常而言，K 值的选取应基于行业经验所得出，尚无明确的理论指导，多数学者所使用的经验规则为 $k_{max}<\sqrt{n}$。本文基于“单一舆情话题所包含的子话题数量不应超过 20 个”这一假设，选取 K 在[2,20]这一范围进行 19 次文本聚类，并对比不同 K 值下的文本聚类有效性系数，从而得出当前最优话题数量 K′。

根据周开乐等的研究[15]，针对聚类的有效性指标可分为 3 类：内部有效性指标、外部有效性指标和相对有效性指标。由于针对网络舆情的文本聚类是无监督学习过程，外部信息是不可用的，内部有效性指标是应用最广泛的聚类有效性指标。而针对内部指标，又通常分为 3 种类型：基于数据集模糊划分的指标、基于数据集几何结构的指标和基于数据集统计信息的指标。基于数据集样本几何结构的指标，根据数据集本身和聚类结果的统计特征对聚类结果进行评估，并根据聚类结果优劣选取最佳聚类数。这一类型的指标包括了 Davies-Bouldin(DB)指标、Calinski-Harabasz(CH)指标、Dunn 指标等。本文采用最为常用的 DB 指标作为文本聚类有效性的判断依据。

DB 指标利用类内样本点到其所属簇集聚类中心的距离来计算类内的紧致性，而用各簇集聚类中心之间的距离来表示类间的分离性，具体定义为：

$$DB(k)=\frac{1}{k}\sum_{i=1}^{K}\max_{j=1\sim k,j\neq i}\left(\frac{W_i+W_j}{C_{ij}}\right)$$

其中，K 为聚类数目，W_i 表示某一类 C_i 中所有样本到该类聚类中心的平均距离，W_j 表示该类 C_i 中所有样本到另一类 C_j 的聚类中心的平均距离，C_{ij} 则代表 C_i 和 C_j 两个类的聚类中心的距离。根据 DB 指标的定义不难发现，该指标越小，则说明类与类之间的相似度越低，从而对应更佳的聚类结果。

通过计算 19 次不同 K 值下的 DB 指标，选取拥有最小 DB 指标的 K 值作为最优话题数量 K′，当 K′ 发生变化时，即可认为原舆情数据中产生了新生的子话题（K′上升），或部分子话题发生了消亡（K′下降）。

3.6 结合时间序列的网络舆情衍进指数

根据不同时刻的最优话题数量 K′，结合舆情本身衍进过程的时间序列，即可得出某一特定时刻 T_i 的网络舆情衍进指数 PODI(public opinion derivative index)，具体定义为：

$$PODI=\frac{K_i'-K_c'}{T_i-T_c}$$

其中，K_i'为 T_i 时刻的最优话题数量，K_c'代表最优话题数量变化为当前最优话题数量前的 K′，T_c 则代表最优话题数量变为 K_i'的时刻。根据 PODI 的定义，由于 K′呈阶梯型变化趋势，因此 PODI 也随之呈现阶梯型变化，并在 K′不发生变化时，呈随时间而下降的趋势。

4 网络舆情衍进指数实证

4.1 数据源选择

本文采用 2017 年 11 月至 12 月间在网络空间引发热议的“教科书式老赖”事件，针对网络舆情衍进指数进行实证研究。2017 年 11 月 22 日，由于被告黄淑芬声称自己没钱，拒绝按照唐山市丰润区人民法院于 2017 年 6 月 8 日的判决结果赔付原告赵香斌 85 万元，原告赵勇在微博上发表了名为“请看什么是教科书式的耍赖”的微博，并曝光了他催促黄淑芬履行法律判决时的对话，从而在全国网络空间中引起极大反应，是当前网络舆情的典型事件，相关舆情量于 2017 年 11 月 23 日达到顶峰后逐渐老化。基于此，本文选取该事件在微博、微信公众号、今日头条 3 个自媒体平台上的相关信息作为数据源，以“教科书式老赖”“黄淑芬”“认真的赵先森”3 个关键词分别进行检索。

4.2 数据采集

本文数据采集的时间窗为 2017 年 11 月 22 日 0:00至 2017 年 12 月 12 日 0:00，从微信公众号、微博、今日头条 3 个自媒体平台上针对“教科书式老赖”突发事件网络舆情相关信息(包括原创、转发及评论)共获取 26 848 条数据。获取的数据库字段包括用户名、用户 ID(UID)、标题、作者、发布时间、发布内容、抓取时间、图片标签、图片内容、视频地址、视频描述等。本文在 3 个自媒体平台上针对“教科书式老赖”事件获取数据的过程如下：①针对微博数据，利用“八爪鱼”数据采集器，对在微博搜索中键入 3 个关键词后的数据进行采集；采集内容包括微博中所有的原创、转发和评论的内容、发表时间与用户信息、所包含图片描述、所包含视频描述信息等。最终共搜集相关原创微博、转发

内容和评论 14 652 条。②针对微信公众号,利用微信自带的"搜一搜"功能,搜索 3 个关键词后显示的公众号内容进行人工采集。采集内容包括公众号名、公众号描述信息、原创文章数量、公众号文章题目、文章作者、发布时间、文章内容、文章阅读数量、文章点赞数量、包含图片描述、包含视频描述、视频地址、评论时间、评论用户名、评论内容等。最终共搜集相关微信公众号文章与评论 8 554 条。③针对今日头条,采用网页端今日头条搜索 3 个关键字,同样进行人工采集;采集内容包括文章标题、作者、文章内容、发布时间、图片描述、评论用户、评论内容、评论发布时间等,最终共搜集相关今日头条文章与评论 3 642 条。其他针对采集数据的描述如表 1 所示:

表 1 采集数据描述信息

来源	类型	数量	平均文本信息量（字）	原创内容平均评论数量(条)	原创内容平均含图片/视频数量(个)
微博	原创内容	2 104	159.71	5.96	0.87
	评论	12 548	22.62		
微信公众号	原创内容	1 317	577.32	5.49	3.42
	评论	7 237	17.86		
今日头条	原创内容	352	1 458.94	9.35	4.51
	评论	3 290	18.24		

4.3 数据处理与分析

在数据处理与分析阶段,本文采用 Excel 整理数据,规范化处理获取的数据字段后,利用 jieba 分词工具对所获得的文本进行切词、删除停用词等操作,形成最小语义单元。此后利用 Python 自带的 IDLE 开发环境对语料库构建词袋空间、计算 TF-IDF 值、利用 K-Means 算法进行文本聚类。随后,利用 DB 指数对文本聚类结果加以检验,最后得出"教科书式老赖"于 2017 年 11 月 22 日至 2017 年 12 月 11 日每一天中的最优话题数量,得出每一天的话题衍进指数。

4.3.1 舆情信息量时间分布 "教科书式老赖"舆情事件的时间演化分布如图 2 所示。在本文所截取的时间范围内,未有明显的舆情潜伏期,但其实该事件早在 2017 年 4 月至 6 月就已开始发酵。2017 年 6 月 8 日,唐山市丰润区人民法院判决被告黄淑芬承担事故主要责任判赔偿 85 万元,故 2017 年 6 月 8 日至 2017 年 11 月 22 日期间可视为此次舆情事件的潜伏期,这一期间网络舆情信息数量相对较少,但持续时间较长。2017 年 11 月 22 日,因黄淑芬一直声称自己没钱拒绝赔偿,同时也拒绝与赵勇沟通,赵勇以"认真的赵先森"这一 ID 在微博发表了"请看什么是教科书式的要赖!"的博文,并曝光了他催促黄淑芬履行法律判决的对话,引爆了舆情演进的热点,舆情迅速进入突发期(2017 年 11 月 22 – 23 日)。从 2017 年 11 月 23 日起舆情演进进入蔓延期(2017 年 11 月 23 日 – 12 月 2 日),三大自媒体平台上用户所发表的原创信息、转发和评论数上升至峰值,最高信息数量达到每日 10 045 条,蔓延期是网络舆情事件数据的主要集中阶段。至 2017 年 12 月 1 日,赵勇父亲赵香斌经抢救无效死亡,至此舆情进入消散期(2017 年 12 月 2 – 11 日),原创舆情信息、评论和转发数均大幅下降。

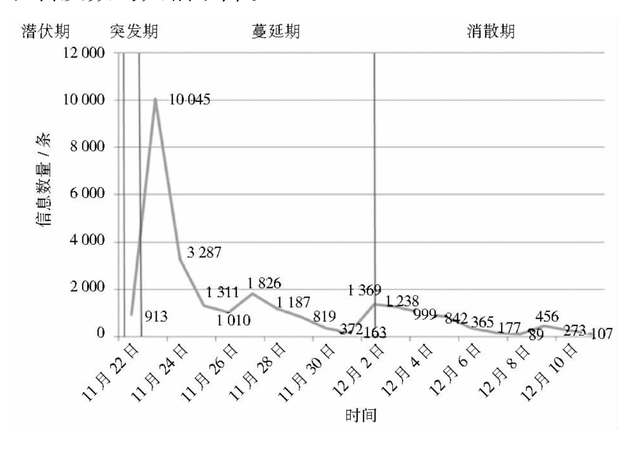


图 2 "教科书式老赖"舆情事件事件时间分布

4.3.2 最优话题数量时间分布 "教科书式老赖"舆情事件最优话题数量的时间演化分布如图 3 所示:

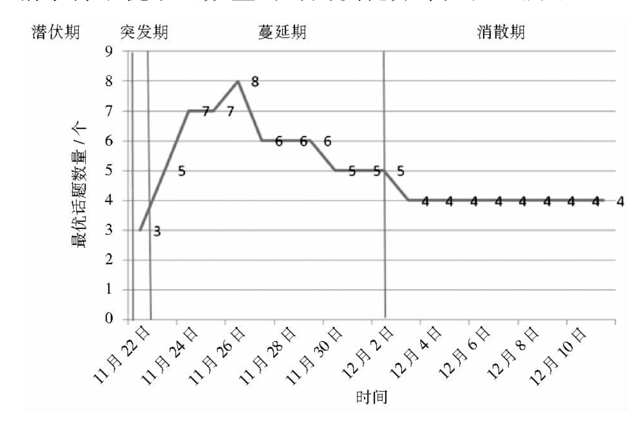


图 3 "教科书式老赖"舆情事件最优话题数量时间分布

由文本聚类结果的 DB 指数可知,在舆情事件进入突发期时,舆情信息仅分为 3 种不同类型的话题,而当舆情进入蔓延期后,舆情开始逐渐发酵,网民针对该事件的讨论逐渐深入、分化,舆情信息逐渐分化,话题数量在 11 月 26 日时达到峰值,达到 8 个类型之多。在进入舆情蔓延期的后半段,随着舆情信息数量的不断减少,话题数量又开始逐渐下降,最终在进入消散期

后的第二天(12 月 3 日),降至 4 个后停止变化。此时可认为,网民针对该事件的讨论方向已经衍进完毕,不再出现变化。在舆情演进不同阶段的 3 个不同时间点上的话题数量以及该话题所对应的文本聚类的关键词如表 2 所示:

表 2　不同阶段最优话题数量与对应关键词

舆情演进阶段	最优话题数量(个)	话题编号	话题对应关键字	话题内涵
突发期	3	1	黄淑芬、刘明月、母女、女儿、无赖、老赖、良心、人渣、买房、愤怒……	对黄淑芬母女行为表达愤怒
		2	赵勇、车祸、医药费、肇事、受害、赔偿、医疗费、官司、司机、正义……	同情支持赵勇
		3	法院、执行、冻结、强制、法律、制裁、强制、执行、迟到、力度……	质疑执法部门不作为
蔓延期	8	1	黄淑芬、刘明月、母女、恶人、无赖、老赖、良心、人渣、买房、耍赖……	对黄淑芬母女行为表达愤怒
		2	赵勇、车祸、医药费、肇事、受害、赔偿、医疗费、善良、司机、官司……	同情支持赵勇
		3	法律、保险、制裁、人肉、拘留、枪毙、逮捕、唐山、和解、耻辱……	为赵勇出谋划策
		4	法院、执行、冻结、正义、强制、暴力、执法、判决、力度、责任……	质疑执法部门不作为
		5	老人、父亲、赵香斌、受害人、安好、抢救、高尚、心疼、祝福、康复……	关注受害人赵香斌
		6	举报、网络、微博、舆论、法律、黑名单、媒体、反思、司法、道德……	反思处理“老赖”事件的有效方式
		7	员工、代理、辞退、公司、工资、高管、领导、开除、财产、解决……	对黄淑芬所在中国平安公司提出要求
		8	平安保险、后悔、不买、形象、蔑视、高层、产品、抵制、态度、企业……	对中国平安公司本身表示愤怒
消散期	4	1	黄淑芬、刘明月、母女、老赖、愤怒、人渣、人性、耍赖、良心、卑鄙……	对黄淑芬母女行为表达愤怒
		2	赵勇、车祸、医药费、肇事、受害、赔偿、医疗费、司机、老人、赵香斌……	同情支持赵勇
		3	微博、网络、法律、保险、制裁、舆论、媒体、道德、举报、司法……	为赵勇出谋划策
		4	法院、执行、冻结、正义、强制、迟到、执法、判决、力度、责任……	质疑执法部门不作为

通过对文本聚类结果的话题关键字的分析,可知“教科书式老赖”事件自始至终最关键的 3 个话题为:对黄淑芬母女表示谴责、同情受害人赵勇和质疑执法部门不作为。在进入蔓延期后,话题衍进至 8 个子话题之多,除上述 3 个话题外,另有为赵勇出谋划策、关注受害人、反思“老赖”事件的有效处理方式、对黄淑芬平安保险公司提出要求、质疑平安保险公司本身等 5 个子话题。最终,舆情进入消散期后,舆情热度逐渐下降,原有的 8 个子话题逐渐消散或由于内容趋同而被并入最终的 4 个话题中,至此话题衍进停止。

4.3.3　话题衍进指数时间分布　根据不同时间点上最优话题数量所得出的话题衍进指数如图 4 所示。由图 4 可知,话题衍进指数在突发期,即舆情迅速演化、话题急剧分化时达到最高。进入蔓延期后,每日新话题产生速度下降,话题衍进指数逐渐下降,并在最后话题数量达到峰值后,转为负值。11 月 27 日,舆情衍进指数由 1 剧烈下降至 -2,这一阶梯式的下降过程说明舆情子话题由增长转为减少,舆情内容本身由发散转为收敛。进入舆情消散期后,最优话题数量趋于稳定,不再发生变化,话题衍进指数稳定在负数,且绝对值不断下降。同时结合图 3、表 2 及图 4,不难发现,“教科书式老赖”舆情事件中,舆情话题就如舆情信息量本身一般,在突发期 - 蔓延期 - 消散期这一过程中,经历了爆发——发展——收敛——稳定的发展过程。

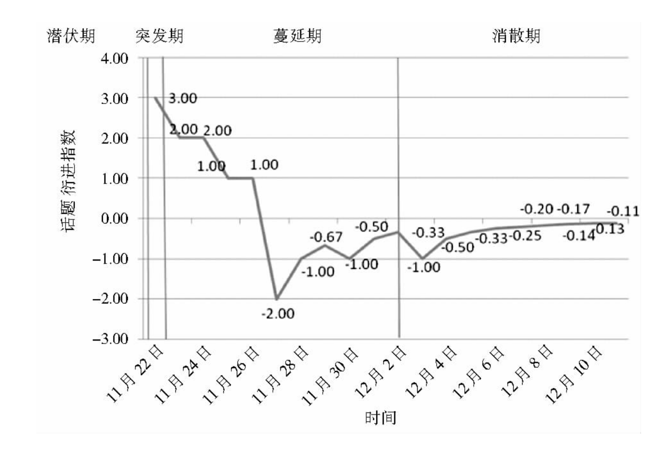


图 4　“教科书式老赖”舆情事件话题衍进指数(PODI)时间分布

4.4　实验结果讨论

以往将聚类算法和舆情分析相结合的研究中,其针对算法本身的有效性及科学性主要通过聚类算法的准确率来进行计算。准确率(Precesion)评价标准的计算公式如下所示:

$$P_{percision} = N_{correct} / N_{total}$$

其中 $N_{correct}$ 指聚类算法正确的文档和实际类别一致的文档数量,N_{total} 为实际聚类的文档数量。

实证分析数据中所包含子话题类别及数量不存在客观的界定方式。为了验证本文提出的根据文本聚类有效性 DB 指数所得出的聚类效果,笔者另采集了

2019年5月内国际政治、美食、教育、医疗4个类别的微博评论信息,共2 240条评论。在实际计算过程中,为多次验证聚类算法的准确率,首先采用国际政治、美食、教育3类微博数据作为数据集1,再采用全部4类微博评论数据作为数据集2。针对数据集1和数据集2,分别用DBSCAN算法、层次聚类算法和本文采用的引入DB指数的改进K-Means算法,从准确率上进行对比,实验结果如表3所示:

表3 文本聚类算法准确率对比

算法		DBSCAN	层次聚类	改进K-means
平均准确率	数据集1	65.3%	69.2%	75.7%
	数据集2	53.9%	63.6%	67.3%

这一部分的实验结果表明相较于传统的DBSCAN和层次聚类算法,引入聚类有效性DB指数的改进K-Means算法可以更加准确地对舆情信息进行分类。故实证分析中各时刻最优子话题数量及聚类的结果较为客观,可视为该时刻下子话题的正确聚类。根据该聚类结果所得出的舆情衍进指数也因此具有了较好的有效性,可以认为在描述舆情衍进这一现象上具备客观性。

5 结论与展望

本文在理论层面,对舆情演进和舆情衍进的概念进行了辨析,并利用文本聚类和文本聚类有效性指数对舆情衍进的具体含义进行了探讨。在实践层面上,以"教科书式老赖"事件作为数据源,通过探讨舆情演进事件、舆情热度、舆情演进不同阶段的文本聚类结果、舆情衍进指数等,以及对比传统的DBSCAN算法和层次聚类算法与本文引入的改进K-Means算法的准确率,证实了舆情衍进指数在描述舆情衍进这一现象时的可行性。同时,根据实证分析中舆情衍进指数的变化可知:舆情在突发期阶段话题衍进速率最高,此后逐渐下降,这一阶段的舆情衍进最为剧烈,子话题的出现呈现爆发性增长;舆情衍进指数在舆情蔓延期内出现阶梯式下降,此后保持为负值,此时舆情的子话题开始逐渐减少,舆情内容本身由发散转为收敛;进入消散期后,子话题数量趋于稳定,舆情衍进指数保持为负值并不断趋近于0。舆情衍进指数作为舆情衍进速率的测度和舆情衍进的判别标准,为舆情监管和舆情预警提供了全新的角度。

下一阶段的研究中,作者将应用舆情衍进指数对不同领域、不同类型的舆情事件进行交叉对比,以确定舆情事件中舆情话题是否都遵循爆发——发展——收敛——稳定这一发展规律。另一研究方向则是结合多媒体识别,将多媒体舆情转为文本后,与文本舆情一同进行文本聚类,对比其与纯文本舆情是否存在差异。

参考文献:

[1] 邓若伊. 论自媒体传播与公共领域的变动[J]. 现代传播(中国传媒大学学报),2011(4):167-168.
[2] 高宾,王兰成. 网络衍生舆情的传播模型及分析方法研究[J]. 情报理论与实践,2019,42(3):166-170,165.
[3] SAITO K, OHARA K, YAMAGISHI Y, et al. Learning diffusion probability based on node attributes in social networks[C]// /International symposium on methodologies for intelligent systems. Warsaw: Springer, 2011:153-162.
[4] WATTS D J, DODDS P S. Influentials, networks, and public opinion formation[J]. Journal of consumer research, 2007, 34(4):441-458.
[5] 兰月新,董希琳,曾润喜,等. 信息异化视角下网络舆情衍生效应模型研究[J]. 情报杂志,2015,34(1):139-143,149.
[6] 陈福集,陈婷. 基于SEIRS传播模型的网络舆情衍生效应研究[J]. 情报杂志,2014,33(2):108-113,160.
[7] 尹熙成,朱恒民,马静,等. 微博舆情话题传播的耦合网络模型——分析话题衍生性特征与用户阅读心理[J]. 情报理论与实践,2015,38(11):82-86.
[8] 王丽君,戴建华. 网络舆情衍生链的定量分析方法研究[J]. 情报科学,2016,34(7):59-63.
[9] 靳晓宏,王强,付宏,等. 主题事件舆情指数的构建及实证研究——以食品安全主题为例[J]. 情报理论与实践,2016,39(12):103-108.
[10] 贺恩锋,庄林远,徐文根. 网络舆情潜在影响力指标体系构建及应用[J]. 情报杂志,2014,33(1):114-119.
[11] 邓尚民,董亚倩. 基于AHP的高校网络舆情安全评估指标体系构建研究[J]. 情报杂志,2012,31(8):31-36.
[12] 陈福集,黄江玲. 我国网络舆情演变文献研究综述[J]. 情报杂志,2013,32(7):54-58,92.
[13] 周昕,黄微,滕广青,等. 网络舆情传播模式解析与重构研究[J]. 情报理论与实践,2016,39(12):25-30.
[14] 黄微,王洁晶,赵江元. 微博舆情信息老化测度研究[J]. 情报资料工作,2017(6):6-11.
[15] 周开乐,杨善林,丁帅,等. 聚类有效性研究综述[J]. 系统工程理论与实践,2014,34(9):2417-2431.

作者贡献说明:

黄微:论文写作指导;
朱镇远:论文选题及撰写;
许烨婧:数据采集及整理;
孙悦:论文文字校对。

黄微，朱镇远，许烨婧，等. 网络舆情衍进指数构建与实证分[J]. 图书情报工作,2019,63(20):26－33.

Establishment of Public Opinion Derivative Index: An Empirical Study in China

Huang Wei Zhu Zhenyuan Xu Yejing Sun Yue
School of Management, Jilin University, Changchun 130022

Abstract: [**Purpose/significance**] During the evolution of public opinion, the derivation of public opinion could possess significant value for the forecasting and warning of public opinion both theoretically and empirically. [**Method/process**] To investigate the mechanism of public opinion derivation, this paper conducted the study using text clustering and DB cluster validity index. It proposed certain standards to judge the occurrence of public opinion derivation and its according velocity index. Furthermore, this paper used an well-known public opinion incident called "Classic Deadbeat" to conduct an empirical research. [**Result/conclusion**] The result of empirical study shows that: the derivative index reaches its climax during emergence phase and declined thereafter. The number of sub-topics reaches its climax during the integration phase and then declined thereafter; when the number of sub-topics decreased, the derivative index become negative, indicating that the public opinion become stabilized. When the public opinion incident reaches the disappearance phase, the number of sub topics become stable and the derivative index remain negative but approach zero. The study of derivative index of public opinion offers a new angle to study public opinion observation and prediction.

Keywords: public opinion derivative derivative index text clustering cluster validity indexes